https://doi.org/10.56946/jspae.v2i1.186
ISSN: 2957-9082
Journal of Soil, Plant and Environment

Review

Exploring Soil Spatial Variability with GIS, Remote Sensing, and Geostatistical Approach

Sangita Singh[1*], Kiranmay Sarma[1]

[1]University School of Environment Management, Guru Gobind Singh Indraprastha University, New Delhi, India
Corresponding Author: sangita.usem.007163@ipu.ac.in

Received: 21 April 2023
Revised: 22 May 2023
Accepted: 03 June 2023

ABSTRACT: This article provides a thorough overview of a wide range of advanced statistical methods that have found extensive and resilient applications in the intricate field of spatial modeling for variables in a geographical information system (GIS) platform. The noteworthy triumph of these approaches can be due to a convergence of speed, dependability, precision, and an inherent eco-consciousness that coexist to reshape the scenario of environmental data analysis. The utilization of these models has outshined conventional methods in the present terrain of scientific investigation and environmental analysis, becoming an authentication of innovative research and decision-making procedures. These approaches demonstrate commendable data utilization efficiency by effectively accepting reduced sample sizes. This not only saves resources but also aligns with the ethical imperative of minimizing environmental effects wherever possible. Furthermore, the combination of these statistical techniques with GIS has paved the way that greatly expands their utility. This tool helps to discover deep spatial linkages, extrapolate trends, and findings into actionable insights that are relatable across all disciplines. These approaches encompass not only predictive modeling but also the realms of error assessment and efficiency evaluation. In conclusion, the adoption of these statistical methods is quite useful in facilitating sound decision-making environmental studies. Some of the domains include soil properties, air quality parameters, vegetation distribution, land cover and land use, water quality parameters, temperature and climate variables, natural hazards, urban infrastructure planning, ecological habitats, noise pollution levels, and radiation and exposure assessment. As the trajectory of scientific growth unfolds, these techniques will serve in directing researchers, practitioners, and policymakers to a future where empirical accuracy and environmental consciousness meet synergistically.

KEYWORDS: Geostatistical, error, prediction, spatial analysis, deterministic, environment.

1. Introduction

Geostatistics, as a specialized branch of statistics, focuses on analyzing, modeling, and interpreting spatial data, providing valuable insights into spatial relationships and the variability of natural phenomena across different locations in a GIS (Geographical Information System) environment. GIS (Openshaw and Clarke, 2019; Wang and Liu, 2023) is fundamentally a framework for gathering, organizing, analyzing, and visualizing diverse types of geographic information. It enables users to grasp detailed relationships, discover patterns, and uncover trends within a geographic environment by

seamlessly integrating data such as maps, satellite imagery, topography data, and attribute data (on-ground analysis data). It is a powerful visualization tool, that translates raw data into understandable maps and graphics by utilizing the knowledge of geography. Initially developed for estimating ore reserves in the mining industry, geostatistics has now found extensive application in diverse fields such as geology (Xu and Zhang, 2023), environmental science (Ghute et al., 2023), agriculture (Mathenge et al., 2022), hydrology (Demarquet et al., 2023), and many more. The process of constructing and analyzing mathematical or statistical representations of spatial relationships, trends, and variations within a geographic area is referred to as spatial modeling. It entails using data to construct models that capture the spatial distribution of phenomena such as environmental factors throughout a specific geographical location. These models seek to elucidate the underlying patterns, relationships and influences that govern the distribution of these occurrences. In the domain of soil science (Khallouf et al., 2020; Criado et al., 2021), geostatistics plays a crucial role in understanding the spatial variability of soil parameters. Soil, being a complex and heterogeneous medium, exhibits significant variations over short distances. The geostatistical analysis aids in characterizing spatial variability (AbdelRahman et al., 2020), creating spatial models (Zakeri and Mariethoz, 2021), and making reliable predictions (Kingsley et al., 2019) of soil properties at unsampled locations. Modern geostatistical tools and techniques, such as semivariograms, spatial auto-correlogram, and various interpolation approaches, are employed to assess the spatial variability (Gökmen et al., 2023; Khan et al., 2021) of soil properties.

In contrast, classical statistical techniques typically rely on descriptive statistical tools like mean, median, mode, coefficient of variation, etc., to measure soil property variability without considering its spatial dependence on the sampling point. However, they fail to adequately explain the continuous spatial variability pattern. Key tools of geostatistics (Gangopadhyay and Reddy, 2022) include variogram, kriging interpolation, spatial uncertainty, and cross-validation. The variogram is a fundamental concept that quantifies the spatial correlation structure in the data by measuring the average difference in values between pairs of data points as a function of their separation distance or lag. It helps determine the range of spatial influence, identify trends, and select appropriate interpolation methods. Variogram models are commonly employed to describe the spatial correlation in the dataset. The interpretation of variograms (Fischer, 2019) involves three components: Sill, Range, and Nugget effect. The Sill represents the plateau or "sill" at large lag distances, signifying the maximum spatial variability. This plateau indicates the range of influence beyond which data points are not significantly correlated. The Range is the distance at which the variogram levels off, indicating the spatial correlation range of the soil parameter, with data points within this range showing a strong correlation. Lastly, the Nugget effect represents the abrupt change in the variogram at a lag distance of zero.

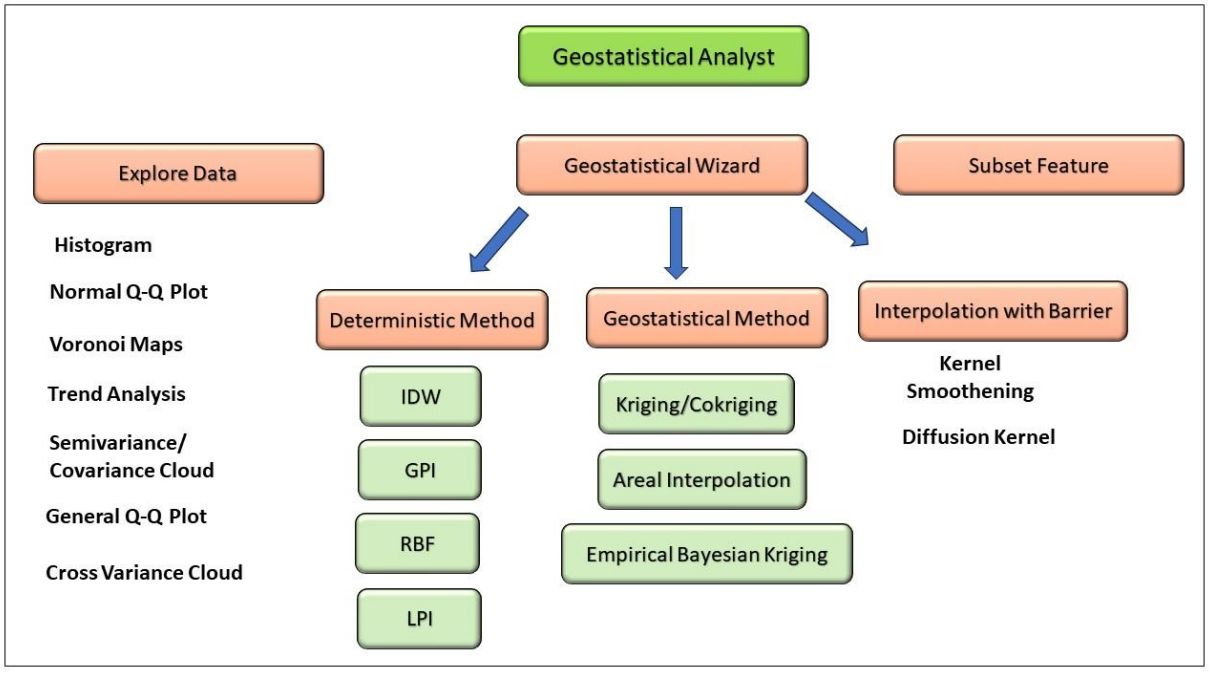

Figure 1: Methods to map spatial variability in parameters by using geostatistical techniques in a GIS platform.

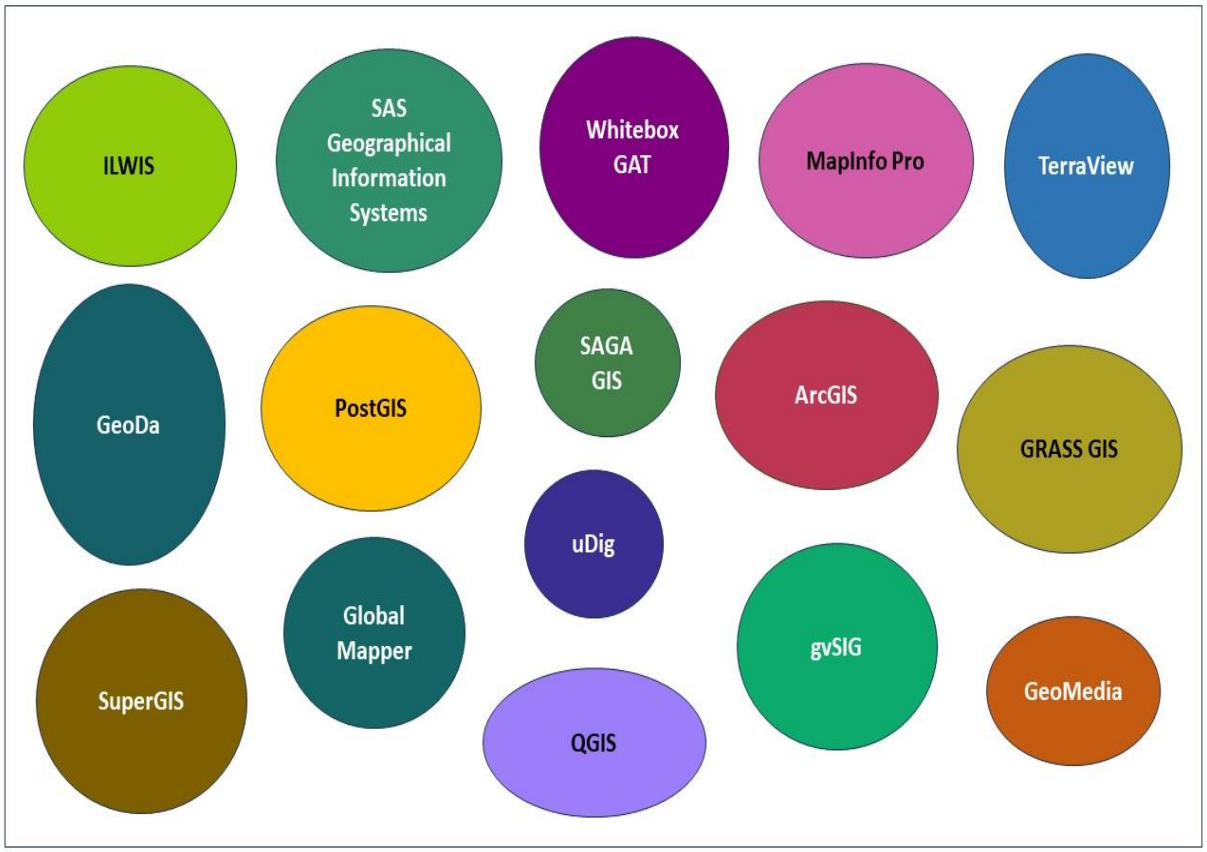

Figure 2: Software to study the spatial datasets.

The variogram plays a crucial role in geostatistics as it accounts for measurement errors, microscale variability, or other factors causing spatial variation at very small distances. It is visualized through a graph that depicts the pattern of semivariance change with varying distances between two sampling points.

Semivariance is calculated by measuring the dispersion of all observation points from a mean or specific value derived from the dataset. It serves to assess spatial continuity or spatial autocorrelation as a function of distance. When the sampling interval between two locations is smaller than the range distance, the variable is considered spatially autocorrelated. Consequently, the spatial variability assessment of that variable becomes significant for its proper management. The N:S ratio provides insight into the degree of spatial dependence of a soil parameter. Different N:S value ranges, such as <0.25, 0.25-0.75, and >0.75, indicate strong, moderate, and weak spatial variability of a particular soil parameter, respectively. Estimating variogram parameters (sill, range, and nugget effect) involves fitting various theoretical models to the experimental variogram. The choice of the model depends on the data and spatial characteristics of the soil parameter being analyzed. Commonly used variogram models (Molla et al., 2023) include the spherical, exponential, Gaussian, and power models. The spherical model is a simple model with a sharp cutoff at the range, resembling a sphere. Conversely, the exponential model is a smoother and continuous model that gradually approaches the sill. These models aid in capturing the spatial correlation structure and are fundamental for accurate predictions and spatial analysis (Mondal et al., 2021) in geostatistics. The Gaussian model shares similarities with the exponential model but exhibits a more gradual increase in spatial correlation. On the other hand, the Power model is specifically useful for data displaying power-law behavior, often employed for variograms with heavy-tailed distributions.

1.1 Error estimation

To identify the most suitable model for a particular soil property, the selection process involves minimizing the error and maximizing the model's efficiency known as the error calculation or cross-validation. The correctness of the spatial model is checked with the error percentages. The "error percentage" provides a quantitative representation of the difference between predicted and observed values at unsampled locations. The predicted values include the model-generated value at a point, whereas the observed values are the recorded values at the location. It is also known as prediction error or estimation error, and it is used to assess the effectiveness of the chosen interpolation strategy. This broadly incorporates mean absolute error (MAE), root mean square Error (RMSE), and mean squared prediction error (MSPE). The "Mean Absolute Error (MAE)" calculates the average of absolute differences between predicted and observed values, providing a measure of usual error magnitude while ignoring directional differences. The root mean square error (RMSE)" on the other hand, encompasses both error magnitude and direction, expressing the square root of the average squared difference between the two

sets of values. Meanwhile, the "mean squared prediction error (MSPE)" focuses on the squared differences between them, emphasizing bigger errors by squaring. These error percentage measurements provide an idea of the efficiency and precision of their interpolation procedures. A smaller error percentage indicates improved prediction accuracy, whereas a higher error percentage indicates a less accurate prediction. Various cross-validation techniques, such as leave-one-out cross-validation or k-fold cross-validation, are commonly employed to validate geostatistical models (Rajalakshimi et al., 2023). In addition to the Gaussian, exponential, and power models, geostatistics offers various other methods and techniques to enhance spatial analysis and prediction. These methods aim to handle diverse data structures and characteristics, providing detailed insights into spatial variability (Nagaraj et al., 2023) and the correlation of soil properties. The whole analysis of datasets is broadly divided into data exploration tools, deterministic method, geostatistical method and interpolation with barrier method which are described below:

2. Data exploration tools

These are the tools used to explore or understand the dataset in detail. On the basis of their utility and properties they can be further subdivided as histogram, normal Q-Q plot, voronoi maps, trend analysis, semivariogram, general Q-Q plot and cross variance cloud.

2.1 Histogram

A histogram is a graphical (Reza et al., 2016; Xu and Zhang, 2023) representation of a dataset's distribution, offering a visual means to comprehend the underlying frequency or probability distribution of numerical data. The dataset is divided into intervals or bins, and the height of each bar in the histogram corresponds to the frequency or count of observations falling within that bin. Histograms prove invaluable in identifying patterns, understanding central tendencies and data spread, detecting outliers, and visualizing the overall shape of the distribution. As such, they are commonly employed in data analysis and exploratory data analysis (EDA) processes.

2.2 A normal quantile-quantile (Q-Q) plot

A quantile-quantile plot, often abbreviated as a normal Q-Q plot or simply a Q-Q plot, is a graphical tool used to assess (Othmani et al., 2023; Wang and Liu, 2023) whether a dataset adheres to a normal distribution. It is achieved by comparing the quantiles of the dataset against the quantiles of a theoretical normal distribution. When the points on the Q-Q plot closely align along a straight line, it indicates that the data is approximately normally distributed. Conversely, deviations from the straight line in a specific pattern suggest the presence of skewness or heavy-tailed characteristics in the data. If the points on the plot exhibit a clear curvature or an "S" shape, it indicates significant non-normality. Q-Q plots are valuable for detecting departures from normality and are commonly employed in statistics, particularly during EDA. They offer visual insights into the data distribution and can aid in selecting appropriate statistical techniques or deciding on data transformations if normality assumptions are necessary for a particular analysis. In summary, Q-Q plots provide a powerful tool for evaluating the conformity

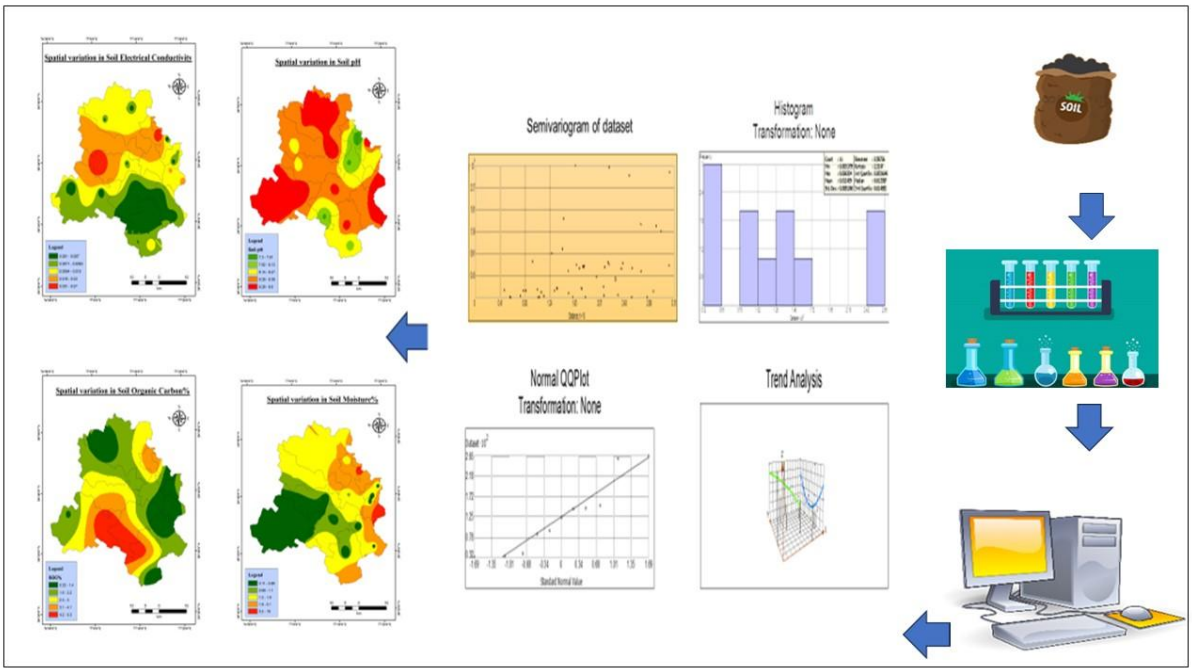

Figure 3: From ground to final output (Source: Singh and Sarma, 2023).

of a dataset to a normal distribution and play a vital role in statistical analyses.

2.3 Voronoi maps

Voronoi diagrams, also known as Voronoi tessellations, are spatial partitions of a given area into regions based on the distance to a set of points called "seeds" or "sites." Each region in a Voronoi map represents the area closest to a particular seed compared to any other seed in the set, and these regions are referred to as Voronoi cells or polygons. Voronoi maps visually depict spatial relationships, illustrating how the study area is divided based on proximity to the seed points. Voronoi maps serve as a powerful tool for understanding spatial (Lu et al., 2022) relationships and find extensive applications in various fields. They efficiently partition space based on distance and are widely used in geography, cartography, spatial analysis, computer graphics, animation, art and design, and many other domains. The versatility of Voronoi diagrams makes them invaluable for analyzing spatial data and visualizing proximity-based patterns in a given area.

2.4 Trend analysis

Trend analysis (Mousavi et al., 2023) is a statistical technique that involves examining the pattern of data over time to identify consistent upward or downward movements, or other patterns, in the data series. This method is widely used in diverse fields, such as economics, finance, marketing, and environmental science (Bangroo et al., 2023), to gain insights into the historical behavior of a variable and make predictions about its future behavior. To quantify the trend in a data set, linear regression or exponential growth/decay models are often utilized. These models help in understanding the direction and magnitude of the trend. Trend analysis serves as a valuable tool to compare present and past trends for a specific variable, allowing for a better understanding of its

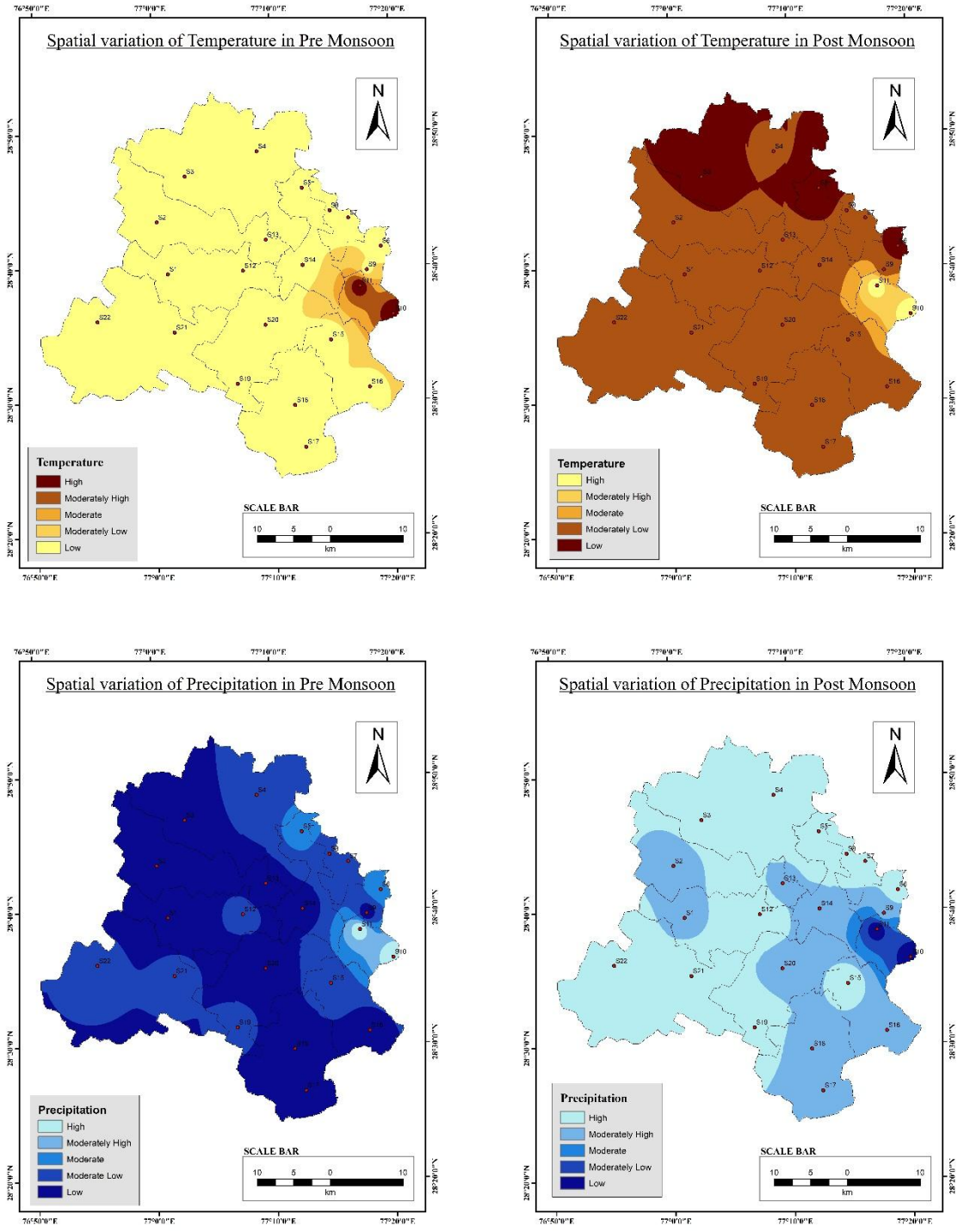

Figure 4: Spatial Maps of Temperature and Precipitation generated by using Kriging

evolution over time. By analyzing historical data trends, decision-makers can make informed predictions and take appropriate actions to respond to changing conditions and plan for the future effectively. Trend analysis enables the identification of important patterns and can provide valuable insights into the underlying factors influencing a particular variable's behavior.

2.5 Semivariogram

The semivariogram, also referred to as a variogram or semivariance function, is a fundamental tool used in geostatistics (Dongare et al., 2022; Fischer, 2019) to analyze spatial variability and quantify spatial autocorrelation within a dataset. This statistical measure illustrates how data points vary concerning their spatial separation or lag distance. Essentially, the semivariogram reveals how the similarity of data values changes with distance. To calculate the semivariogram, one employs semivariance, which is half the average squared difference between data points within a given lag distance. By doing so, it quantifies the level of similarity or dissimilarity between data points at a specific distance apart. When the lag distance is small, the semivariance tends to be low since nearby points exhibit higher similarity. However, as the lag distance increases, the semivariance may increase up to a certain point, representing the spatial autocorrelation range or "nugget." Beyond the nugget, the semivariance may reach a plateau, indicating that the spatial dependence has reached its maximum.

The shape of the semivariogram assists statisticians in identifying the appropriate spatial model for interpolation or prediction. Common models used to fit the semivariogram include the exponential, spherical, and Gaussian models. The estimation of the semivariogram can also be visualized through a covariance cloud (Openshaw and Clark, 2019), which represents the covariance between two variables. Each point in the cloud corresponds to a pair of data points, and its position on the graph reflects their joint covariance. This representation provides further insights into spatial relationships and helps in understanding the spatial structure of the dataset.

2.6 A general quantile-quantile (Q-Q) plot

A general quantile-quantile (Q-Q) plot is a graphical tool used to evaluate whether a dataset adheres to a particular probability distribution. Unlike the normal Q-Q plot, which specifically checks for normal distribution, the general Q-Q plot can be employed to assess the fit of data to various theoretical distributions. In a general Q-Q plot, if the dataset follows the target distribution, the points on the plot will approximately align along a straight line. Deviations from this straight line indicate a difference from the specified distribution. If the points closely follow the straight line, it suggests that the data is well-described by the chosen theoretical distribution. On the other hand, if the points deviate from the line in a systematic pattern, it indicates that the data differs from the target distribution. General Q-Q plots are invaluable tools in statistical analysis, as they provide a visual means to assess the goodness of fit between data and different theoretical distributions. They are particularly useful when determining the most appropriate distribution for modeling the data or when testing assumptions in statistical

methods that rely on specific distributions. By employing general Q-Q plots, researchers can gain insights into the suitability of various distributions for representing their data and make informed decisions about the choice of statistical models (Lu et al., 2022; Nagaraj et al., 2023) and assumptions.

2.7 Cross Variance

Cross variance (Othmani et al., 2023, Fischer, 2019) refers to the covariance between two variables in a multivariate setting. It is a measure of how two variables vary together, capturing the degree of correlation or relationship between them. The cross variance is commonly used in statistical analysis to understand the association between two variables and to assess their joint behavior. To calculate cross variance, one can observe the cross-variance cloud, which is a visual representation of the covariance or correlation between the two variables. This is achieved by plotting the data points of both variables on a scatter plot, with one variable on the x-axis and the other on the y-axis. The resulting cloud of points provides insights into the strength and direction of their relationship. A straight linear-shaped cloud indicates a strong positive or negative correlation between the variables, while a scattered or elliptical cloud suggests a weaker or no correlation. By analyzing (Reza at al., 2016) the cross-variance cloud, researchers can quickly assess the level of association between the two variables and make informed decisions about their relationship in the dataset.

3. The deterministic way

A deterministic method is an algorithm (Molla et al., 2023; Zakeri and Mariethoz, 2021) or approach that consistently generates the same output for a given input, regardless of the number of times it is executed. It operates without any randomness or uncertainty, resulting in a completely predictable and consistent outcome. This quality makes deterministic methods highly valuable in fields such as computer science, mathematics, physics, and engineering, where repeatability, reliability, and precision are essential. Their key characteristics include repeatability, predictability, the elimination of uncertainty, and overall consistency. By offering reproducibility and stability, deterministic methods ensure reliable and accurate results in various applications, ranging from simulations to critical decision-making processes. The deterministic method includes four subdivisions:

3.1 Inverse distance weighting (IDW)

Inverse distance weighting (IDW) is a widely used interpolation technique in spatial analysis and geostatistics for estimating values at unsampled locations based on nearby sampled data points. The fundamental assumption of IDW (AbdelRahman et al., 2020, Openshaw and Clark, 2019) is that values at unsampled locations are influenced more by the values of nearby points than those farther away. To achieve this, the method employs a power parameter "p" that controls the influence of nearby points on the estimation. Typically, "p" is set between 1 and 3, with lower values giving more weight to points closer to the target location and higher values providing more equal weight to all points. IDW is favored for its simplicity and intuitive nature, making it straightforward to implement. However, it does have some limitations. One such limitation is its sensitivity to the choice of the

power parameter, which can affect the interpolation results significantly. Additionally, IDW tends to produce "bull's-eye" artifacts around data points, particularly when the data is sparse or unevenly distributed. As a result, IDW is often utilized for basic interpolation tasks and serves as a baseline for more sophisticated interpolation methods in GIS and spatial data analysis. These advanced techniques (Molla et al., 2023) takes into account additional factors, such as spatial autocorrelation, spatial trends, and variogram models, to achieve more accurate and robust interpolation results for complex spatial datasets. Despite its limitations, IDW remains a valuable tool in geospatial analysis, providing a quick and straightforward solution for certain interpolation needs.

3.2 Global polynomial interpolation (GPI)

Global polynomial interpolation (GPI) is an interpolation technique utilized to estimate values between known data points by fitting a polynomial function to the entire dataset. Unlike local interpolation methods, such as inverse distance weighting, GPI considers the entire dataset to create a single polynomial function that smoothly fits all the given data points. The objective of global polynomial interpolation is to find a polynomial function that accurately passes through all the provided data points, allowing for the approximation of values at any point within the dataset's range. While global polynomial interpolation offers advantages, such as producing a smooth global approximation of the entire dataset, it may not be suitable for datasets (Zakeri and Mariethoz, 2021) with a high degree of noise or outliers. High-degree polynomials can lead to oscillations and overfitting, where the interpolation function becomes overly sensitive to individual data points. The complexity and computational intensity of the interpolation process increase with the degree of the polynomial used. Therefore, selecting an appropriate polynomial degree becomes crucial in balancing the need to capture the data's essential behavior while avoiding overfitting. In practice, when more flexibility and robustness are required in spatial data interpolation, other methods such as spline interpolation or kriging are commonly employed. These techniques provide more adaptive and smoother interpolations, making them suitable for datasets with noise or outliers. By considering the specific characteristics of the dataset, researchers can choose the most appropriate interpolation method to achieve an accurate and reliable estimation of values between data points.

3.3 Radial basis function (RBF)

Radial Basis Function (RBF) is a versatile mathematical function widely utilized for interpolation, approximation, and smoothing of data, especially in scenarios involving scattered data points in multidimensional space. The core concept behind RBF (Singh and Sarma, 2023; Othmani et al., 2023) involves approximating a complex function by a combination of simple functions known as basis functions, which exhibit decaying behavior with distance from the center point. The popularity of radial basis functions lies in their flexibility and adaptability to complex and irregular data patterns. These functions offer smooth and continuous interpolation, even in high-dimensional spaces, making them well-suited for various applications. RBF interpolation finds extensive use in data

smoothing, image processing, computer graphics, and numerical solutions of partial differential equations. By employing RBF, researchers can effectively handle problems with scattered data and achieve accurate estimations (Molla et al., 2023) and approximations in multidimensional space. The ability of radial basis functions to capture intricate data relationships and provide seamless interpolation makes them a valuable tool in data analysis and various computational fields.

3.4 Local polynomial interpolation (LPI)

Local polynomial interpolation (LPI) is an interpolation method (Fischer, 2019) designed to estimate values between known data points by fitting a polynomial function to a small subset of nearby data points. Unlike global polynomial interpolation (GPI), which considers the entire dataset to create a single polynomial function, LPI adapts the interpolation model for each target point based on its neighboring data points. In LPI, the key idea is to construct a polynomial function that better approximates the data around each target point by using a weighted average of nearby data points. The polynomial is usually of a low degree, such as linear or quadratic, to ensure smoothness and prevent overfitting. Each data point in the neighborhood of the target point is assigned a weight based on its distance from the target point. A kernel function is typically employed for this weighting, giving more weight to points closer to the target and less weight to points farther away. For each target point, a polynomial function, such as linear or quadratic, is fitted to the weighted subset of nearby data points using weighted least squares or other regression techniques. This approach (Wang and Liu, 2023) allows LPI to adapt to the changing data behavior more effectively, making it particularly useful for datasets with spatial or temporal heterogeneity, where the underlying data pattern varies across different regions or periods. Local polynomial interpolation finds widespread application in spatial data analysis, geostatistics, and time series analysis, where capturing local variations is crucial for accurate predictions and interpolation. However, the choice of the bandwidth parameter (defining the size of the neighborhood) and the degree of the local polynomial can significantly impact the quality of the interpolation. Hence, careful selection of these parameters is essential to achieving reliable and precise interpolation results.

4. The geostatistical way

Geostatistical methods (Othmani et al., 2023; Rajalakshimi et al., 2023; Lu et al., 2022; Gangopadhyay and Reddy, 2022) comprise a set of statistical techniques specifically designed to analyze and model spatially correlated data. These methods hold significant value for applications in geology, environmental science, mining, agriculture, and other fields where data is collected across different geographic locations. Geostatistics takes into account the spatial dependence or autocorrelation that may exist in the data, enabling more accurate predictions and interpolation of values at unsampled locations. The primary objectives of geostatistical methods are to create spatial maps, identify spatial patterns, estimate values at unsampled locations, and quantify uncertainty in predictions. By considering the spatial relationships between data points,

geostatistical approaches deliver robust and reliable results for decision-making in various fields that heavily rely on spatial data analysis and prediction. Key geostatistical techniques (Dongare et al., 2022; Gangopadhyay and Reddy, 2022; Khan et al., 2021) include spatial autocorrelation, variogram analysis, kriging (including ordinary kriging and universal kriging), co-kriging, and geostatistical simulation. These methods play crucial roles in understanding spatial patterns, predicting unknown values, and managing spatially distributed resources effectively. Overall, geostatistical methods provide powerful tools for handling spatial data, enabling data-driven decision-making, and facilitating informed actions in diverse fields where spatial analysis and prediction are vital. This comprises kriging, co-kriging, areal interpolation and empirical bayesian kriging.

4.1 Kriging

Kriging is a powerful geostatistical (Reza et al., 2016; Rajalakshimi et al., 2023; Mondal et al., 2021) interpolation method that delivers the best linear unbiased estimate of a variable at unsampled locations (Singh and Sarma, 2020). It incorporates both spatial correlation and uncertainty in the data, making it a robust and reliable interpolation technique. The fundamental principle underlying Kriging (Khan et al., 2021; Kingsley et al., 2019; Openshaw and Clark, 2019) is to minimize prediction error by assigning appropriate weights to neighboring data points based on their spatial distance and correlation. The Kriging method assumes that the spatial correlation in the data can be modeled using a variogram (or semivariogram). This variogram describes how the variance of the variable changes with the distance between data points. By using the variogram model, Kriging can provide a continuous and spatially smooth surface, allowing for accurate estimation at unsampled locations. A notable advantage of Kriging is its ability to quantify the uncertainty in predictions. The method produces an estimation variance that indicates the level of uncertainty associated with the estimated values, providing valuable insights into the reliability of the predictions. Various variants of Kriging, including ordinary Kriging, simple Kriging, and universal Kriging, offer different levels of assumptions and complexity. Among these, ordinary Kriging is the most widely used approach for generating spatial variability maps of soil properties due to its superior performance compared to other approaches. In summary, Kriging is a highly effective geostatistical method that accounts for spatial correlation and uncertainty, enabling precise and reliable estimation of values at unsampled locations. Its ability to generate smooth surfaces and provide uncertainty measures makes it a popular choice for various applications in geology, environmental science, agriculture, and more.

4.2 Co-Kriging

Co-Kriging is an extension of the traditional Kriging method used for the simultaneous interpolation of two or more correlated variables. It is particularly beneficial when there is a spatial correlation between multiple variables, and utilizing this correlation can enhance the accuracy of the estimates. Co-Kriging becomes valuable in situations where two or more variables exhibit spatial relationships, and it can provide more precise predictions compared to using Kriging independently, especially when

data for one variable is sparse or missing. Both Kriging and Co-Kriging are powerful tools for spatial interpolation, enabling the estimation of values at unsampled locations while taking into account spatial correlation and uncertainty.

The choice between Kriging (Singh and Sarma, 2020) and Co-Kriging depends on the characteristics of the data and the presence of multiple correlated variables. When multiple correlated variables are available, Co-Kriging can leverage the spatial relationship between them to improve the interpolation results. In summary, Co-Kriging is a valuable geostatistical technique that extends the capabilities of traditional Kriging by allowing for the joint estimation of multiple correlated variables.

4.3 Areal interpolation

Areal interpolation, also known as areal weighting or areal disaggregation, is a technique used to estimate and redistribute data from one set of areal units to another set of non-overlapping areal units. The purpose of areal interpolation is to harmonize spatial data (Dongare et al., 2022; Bangroo et al., 2023; Reza et al., 2016) that are available at different geographic resolutions or administrative boundaries. When using areal interpolation, data representing an entire geographic area (such as a country, region, or municipality) is transferred to a different set of geographic units, which may have different shapes and sizes. The method involves redistributing the data based on some proportional relationship between the areas of the source and target units. Areal interpolation methods can vary depending on the assumptions made about the spatial relationship between the source and target units. Common approaches include the areal weighting method, which redistributes data based on the proportional overlap of source and target areas, dasymetric mapping, which considers additional ancillary data to refine the interpolation, and spatial interpolation techniques like Inverse Distance Weighting (IDW), which use the spatial proximity of data points for redistribution. Areal interpolation is essential for various applications (Zakeri and Mariethoz, 2021), such as harmonizing data from different sources, aggregating data to a common geographic scale, and generating consistent spatial datasets for analysis and modeling across different geographic units. It finds extensive use in fields such as geography, demography, environmental science, and regional planning, where harmonizing and integrating spatial data from diverse sources is crucial for accurate analysis and decision-making.

4.4 Empirical Bayesian kriging (EBK)

Empirical Bayesian kriging (EBK) is an advanced geostatistical interpolation method that combines the principles of kriging and Bayesian statistics to estimate values at unsampled locations. This technique is an extension of traditional kriging and offers several advantages by incorporating external information about the spatial variability of the data. In traditional kriging, the variogram model, which measures spatial correlation, is assumed to be known or directly estimated from the data. However, in empirical Bayesian kriging, the variogram model parameters are treated as random variables and estimated using additional data called the "drift" or "external drift" data. This approach provides greater flexibility in modeling

spatial correlation since variogram parameters are estimated instead of assumed. Empirical Bayesian kriging delivers more reliable uncertainty estimates by considering variogram uncertainty in the interpolation process. It can handle situations where the spatial variability of the data varies across different regions, making it adaptable to complex datasets. However, empirical Bayesian kriging requires the availability of external drift data, which may not always be readily obtainable, and it may be more computationally intensive compared to traditional kriging. Empirical Bayesian kriging finds common application in geostatistics (Lu et al., 2022), spatial data analysis, and environmental modeling (Mondal et al., 2021), particularly when auxiliary information is available to enhance interpolation accuracy and uncertainty estimates. This method is particularly useful for large datasets or situations where data are collected at different spatial scales, providing a powerful tool for spatial data analysis and prediction in various fields.

5. Method of Interpolation with Barriers

The Method of Interpolation with Barriers, also known as constrained interpolation, is a spatial interpolation technique that takes barriers or constraints into account during the interpolation process. Barriers refer to areas in a geographic space where data values are not continuous or where the underlying phenomenon being interpolated is interrupted or discontinuous. The primary objective of interpolation with barriers is to generate a smooth and continuous surface while respecting the presence of barriers and avoiding interpolation across them. This is particularly important in situations where the data or phenomenon being interpolated should not be assumed to be continuous over certain regions. Interpolation with barriers becomes especially useful when certain geographic features act as physical boundaries, such as rivers (Lu et al., 2022), mountains, or land use boundaries (Nagaraj et al., 2023; Othmani et al., 2023). It is essential to consider these barriers when estimating values at unsampled locations to ensure accurate and realistic results. This method finds application in various fields, including environmental modeling, hydrology, urban planning, and natural resource management. By incorporating spatial constraints, interpolation with barriers helps create more reliable and accurate interpolation results, avoiding unrealistic interpolation across physical barriers and providing a better representation of the underlying spatial patterns and variations.

5.1 Kernel smoothing

Kernel smoothing, also known as kernel regression (AbdelRahman et al., 2020; Zakeri and Mariethoz, 2021) or kernel density estimation, is a non-parametric statistical technique widely used to estimate the underlying smooth pattern of a dataset. This method is commonly applied in data analysis and visualization to reduce noise, reveal underlying trends, and estimate probability density functions for continuous data. The fundamental concept behind kernel smoothing is to approximate the values of a function at a specific point by averaging the observed data points, weighted by their distance to that point. The weights are determined by a kernel function, which is a symmetric, non-negative function that decreases as the distance from the point of

interest increases. The kernel function acts as a smoothing window, and its choice impacts the smoothness of the resulting estimate. Kernel smoothing finds applications in various fields, including signal processing, geostatistics, image processing, and environmental science. It is particularly advantageous because it does not assume a specific parametric model, making it flexible and versatile for handling complex patterns and noisy data. However, selecting an appropriate kernel function and bandwidth parameter is crucial to obtain meaningful and accurate results in kernel smoothing. The bandwidth parameter controls the width of the smoothing window and influences the level of smoothing applied to the data. Cross-validation (Gökmen et al., 2023; Kingsley et al., 2019; Bangroo et al., 2023) techniques are often used to optimize the bandwidth selection for a specific dataset and application, ensuring that the resulting estimates are reliable and well-suited to the data characteristics. In conclusion, kernel smoothing is a powerful non-parametric technique that allows for the estimation of smooth patterns in data without making strong assumptions about the underlying model. Its flexibility and versatility make it a valuable tool in various fields for data analysis, noise reduction, and probability density estimation. Properly selecting the kernel function and bandwidth parameter is essential to ensuring the accuracy and meaningfulness of the smoothed estimates.

5.2 The diffusion kernel

The diffusion kernel, also known as the heat kernel or Gaussian kernel, is a specific type of kernel function employed in various mathematical and computational methods, such as machine learning, graph theory, and image processing. Its name originates from its connection to the heat equation in physics, where it represents the diffusion of heat over time. The diffusion kernel is constructed based on a similarity matrix derived from the dataset using techniques like the Gaussian similarity function. This similarity matrix measures the similarity or closeness between data points. The diffusion kernel finds application in diverse areas, including graph-based machine learning, dimensionality reduction, image processing, and spectral clustering. It serves as a potent and versatile tool for capturing the intrinsic structure and relationships in complex datasets (Zakeri and Mariethoz, 2021) and graphs. One of the key advantages of the diffusion kernel is its ability to handle both local and global information diffusion, making it valuable for tasks where understanding relationships between data points at different scales is crucial. Overall, the diffusion kernel is a powerful and flexible approach widely used in data analysis and machine learning tasks due to its capacity to capture complex relationships and patterns within datasets and graphs. Its association with the heat equation adds to its significance in various fields, making it an essential tool in diverse applications.

6. Limitations and challenges

The spatial analysis also comes with certain limitations which preferably include:

Limited data: For reliable predictions, geostatistical approaches require a significant number of soil samples. Data that is sparse or inadequately dispersed might lead to uncertainty and less reliable outcomes.

Data accuracy and quality: The accuracy

and quality of soil samples might have an impact on the dependability of geostatistical analysis. It is critical to ensure data quality through adequate sampling practices and data validation.

Outliers: Outlier data points can have a substantial impact on the findings of variogram modeling and interpolation. Outliers must be identified and handled correctly to ensure the robustness of the analysis.

Assumptions of stationarity: Geostatistical techniques make the assumption that the spatial dependence of soil characteristics remains constant over the whole study region. However, the spatial correlation may vary across the region, putting doubt on the hypothesis of stationarity.

Model selection: Selecting appropriate variogram models and interpolation techniques can be a difficult process. The selection of models and approaches has a considerable impact on the accuracy of predictions and interpretations.

Extrapolation: Making predictions outside of the range of the collected data might be erroneous and result in inaccurate conclusions. Predictions should be used with caution to avoid overestimation.

Prediction uncertainty: Geostatistical forecasts are subject to intrinsic uncertainty. Understanding and quantifying uncertainty is critical for making informed prediction-based decisions.

Scale and resolution: The scale of the study affects geostatistical analysis. The resolution of spatial data might affect the conclusions, and findings at one scale may not hold true at another.

Anisotropy: Anisotropy occurs when the spatial structure of soil characteristics exhibits distinct patterns in various directions. Including anisotropy in variogram modeling can be difficult and may necessitate additional considerations.

Spatial bias: Spatial data may suffer from sampling bias, in which specific portions of the dataset are overfit or underfit. The representativeness of the forecasts can be influenced by spatial bias.

7. Study outlook

Several possible breakthroughs and developments (Parker, 2023) in the field of spatial modeling for environmental factors are likely to alter the way of comprehension and analysis of spatial data. Future developments will most likely focus on incorporating temporal aspects into spatial models, allowing researchers to investigate how environmental variables change over time. Remote sensing technological advancements are enabling the collection of high-resolution and multi-spectral data. This allows for more accurate and realistic depictions of environmental parameters such as land cover, temperature, and vegetation.

Machine learning and deep learning algorithms can extract patterns from this data and use them to create more accurate spatial models. Combining various types of data and modeling approaches, such as simulation models and observational data, will result in a more comprehensive understanding of environmental processes. Hybrid models will require improved statistical methodologies to successfully blend varied data sources. Machine learning algorithms (Mousavi et al., 2023) used with spatial data will provide more accurate and interpretable models.

Spatial convolutional neural networks (CNNs) and graph-based machine learning techniques will be utilized to utilize complex spatial patterns in environmental data.

8. Future perception

Geostatistical methods are more environmentally friendly than traditional approaches as they optimize resource utilization with minimum waste generation, reduce the environmental effect, and deliver more accurate solutions. Geostatistical approaches can uncover geographical patterns (Parker, 2023) and trends in environmental data, allowing for early detection of environmental changes or abnormalities. This enables immediate mitigation efforts, reducing potential harm to ecosystems and minimizing long-term consequences. It can also help discover appropriate locations (site suitability studies) for renewable energy installations such as solar panels and wind turbines, maximizing energy production while minimizing environmental impact. It also helps to improve climate modeling accuracy by including spatial variability (Wang, 2023). This enables more accurate predictions of climate-related variables, which aids in assessing the implications of climate change and guiding adaptation efforts. Generally, traditional approaches may yield accurate answers without the added complexity of spatial modeling in circumstances where data is abundant and spatial patterns are not prominent. Spatial models, on the other hand, are likely to exceed traditional methods in terms of accuracy when dealing with environmental data that exhibit spatial autocorrelation, local variability, or irregular patterns. However, spatial models are used in environmental monitoring and remediation to monitor air and water quality, predict pollutant dispersion, and locate contaminated areas. They direct effective site remediation procedures that reduce exposure hazards and environmental damage. Studies have reported that the adoption of spatial models has led to improvements in real-world scenarios. Spatial models have been used in climate change studies, seasonal changes mapping (Zhao et al., 2018), epidemiology (Kieu et al., 2021) and disease mapping (Abd El-Ghany et al., 2020), high-risk pollution zones (Jumaah et al., 2023), deforestation patterns (Coetzee, 2022), identifying regions at risk of soil erosion, and guiding conservation activities (Ghute et al., 2023), soil nutrient levels, urban planning infrastructure development and resource utilization (Singh and Sarma, 2023). They are well used for agricultural needs whether related to fertilizer or irrigation by minimizing waste and production upsurge. Spatial models have also estimated the scope of natural disasters (Khan et al., 2023) such as floods, landslides, and tsunamis. These forecasts help early warning systems and emergency response preparation. They are also used in energy sector planning, healthcare planning, and various conservation methods.

9. Conclusions

Geostatistics has emerged as an indispensable and versatile tool for understanding spatial relationships and making accurate predictions in various scientific fields. In the context of soil parameter analysis, geostatistics provides a powerful means to comprehend the spatial variability of soil properties.

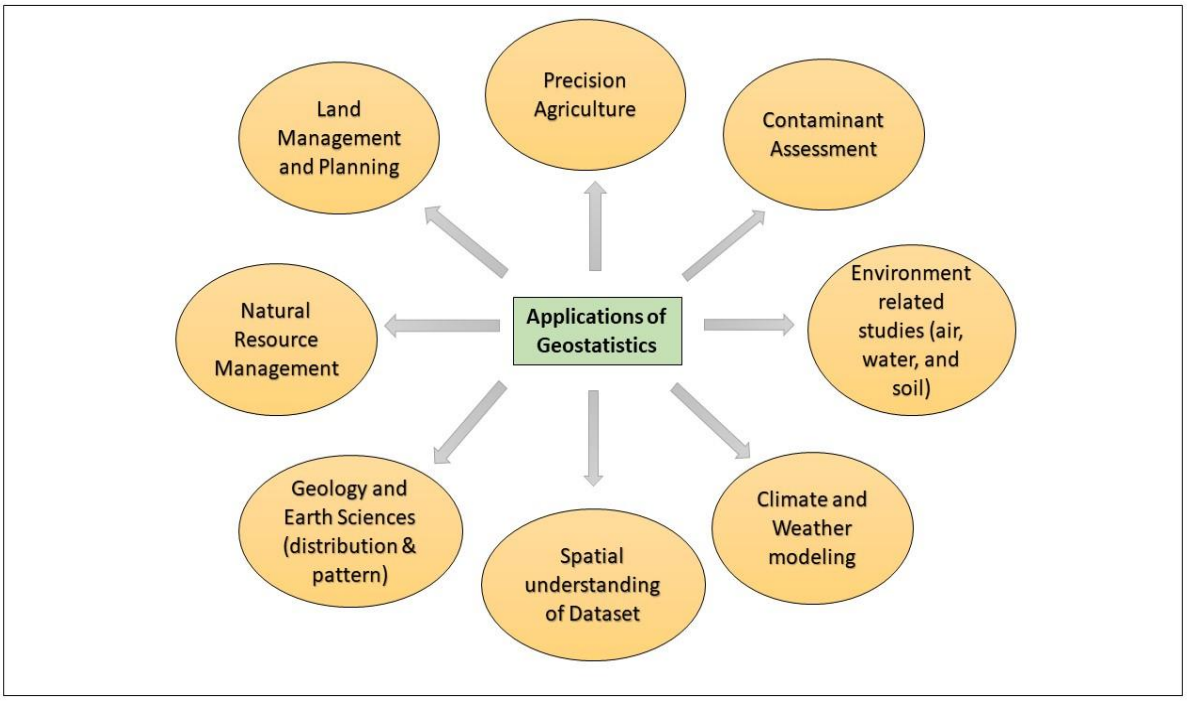

Figure 5: Basic Applications of Geostatistical Analysis

By identifying spatial patterns and quantifying uncertainties, geostatistics facilitates informed decision-making for sustainable land management and environmental applications. The key to harnessing the full potential of geostatistics lies in careful data collection, preprocessing, and validation. Ensuring the accuracy and reliability of results requires diligent attention to data quality and representative sampling techniques. By capitalizing on the spatial autocorrelation inherent in natural phenomena, geostatistical methods offer valuable insights that enable effective management of natural resources, environmental protection, and informed decision-making across a broad spectrum of applications. Nevertheless, it is imperative to approach geostatistical analyses with caution, select appropriate models, and account for uncertainties to ensure robust and trustworthy results. As technology continues to advance and data collection techniques improve, geostatistics will continue to play a vital role in unraveling the complexities of soil systems and promoting sustainable land management practices. In conclusion, geostatistics has revolutionized spatial data analysis, and its application in soil parameter analysis holds immense promise for enhancing our understanding of soil variability and supporting evidence-based decision-making for a sustainable and resilient future.

Acknowledgments The authors would like to thank UGC (F1- 17.1/2015-16/RGNF-2015-17-SC-DEL-27579) for the financial assistance provided to the first author. The authors would like to extend their gratitude to the anonymous reviewers who have

contributed to the improvement of the manuscript and a huge appreciation for providing the necessary facilities goes to the Department of Environment Management, Guru Gobind Singh Indraprastha University, New Delhi (India).

Conflicts of Interest The authors declare no conflicts of interest.

Availability of Data and Materials Data will be available on a formal request from the corresponding authors.

Funding UGC (Rajiv Gandhi National Fellowship) fellowship to the first author (Awarded: F1-17.1/2015-16/RGNF-2015-17-SC-DEL-27579).

Authors Contributions S.S.—writing, conceptualization, investigation, editing, formal analysis, and drafting study; K.S.—review, supervision.

REFERENCES

Abd El-Ghany, N. M. Abd El-Aziz, S. E. and Marei, S. S. 2020. A review: application of remote sensing as a promising strategy for insect pests and diseases management. Environmental Science and Pollution Research. (2020). 27.33503-33515.

AbdelRahman, M. A. Zakarya, Y. M. Metwaly, M. M. and Koubouris, G. Deciphering soil spatial variability through geostatistics and interpolation techniques. Sustainability. (2020).13(1).194.

Bangroo, S. A. Bhat, M. I. Wani, J. A. Rasool, R. Madhi, S. S. Bashir, O. and Shah, T. I. Mapping soil properties using geostatistical methods for mid to high altitude temperate zone of Kashmir Himalayas. Journal of the Indian Society of Soil Science. (2023). 71(1).1-12.

Coetzee, C. Change Detection of Vegetation Cover Using Remote Sensing and GIS–A Case Study of the West Coast Region of South Africa. Geography, Environment, Sustainability. (2022). 15(2).91-102.

Criado, M. Martínez-Graña, A. Santos-Francés, F. and Merchán, L. Improving the Management of a Semi-Arid Agricultural Ecosystem through Digital Mapping of Soil Properties: The Case of Salamanca (Spain). Agronomy. (2021). 11(6).1189.

Demarquet, Q. Rapinel, S. Dufour, S. and Hubert-Moy, L. Long-Term Wetland Monitoring Using the Landsat Archive: A Review. Remote Sensing. (2023). 15(3).820.

Dongare, V. T. Reddy, G. P. Kharche, V. K. and Ramteke, I. K. Spatial variability of soil nutrients under sugarcane cropping system in semi-arid tropics of western India using geostatistics and GIS. Journal of Soil and Water Conservation. (2022). 21(1).67-75.

Fischer, M. M. (2019): Spatial analytical perspectives on GIS. Routledge, London.

Gangopadhyay, S.K. and Reddy, G.P. Spatial variability of soil nutrients under the rice-fallow system of eastern India using geostatistics and Geographic Information System. Journal of Soil and Water Conservation. (2022). 21(1).55-66.

Ghute, B. B. Shaikh, M. B. and Halder, B. Impact assessment of natural and anthropogenic activities using remote sensing and GIS techniques in the Upper Purna River basin, Maharashtra, India. Modeling Earth Systems and Environment. (2023). 9(2).1507-1522.

Gökmen, V. Sürücü, A. Budak, M. and Bilgili, A. V. Spatial modeling and mapping of the spatial variability of soil micronutrients in the Tigris basin. Journal of King Saud University- Science. (2023). 102724.

Jumaah, H. J. Ameen, M. H. Mahmood, S. and Jumaah, S. J. Study of air contamination in Iraq using remotely sensed Data and GIS. Geocarto International. (2023). 38(1).2178518.

Khallouf, A. AlHinawi, S. AlMesber, W. Shamsham, S. and Idries, Y. Digital Mapping of Soil Properties in the Western-Facing Slope of Jabal Al-Arab at Suwaydaa Governorate, Syria. Jordan Journal of Earth and Environmental Sciences. (2020). 11(3).193-201.

Khan, M. Z. Islam, M. R. Salam, A. B. A. and Ray, T. Spatial variability and geostatistical analysis of soil properties in the diversified cropping regions of Bangladesh using geographic information system techniques. Applied and Environmental Soil Science. (2021). 1-19.

Khan, S. M. Shafi, I. Butt, W. H. Diez, I. D. L.T. Flores, M. A. L. Galán, J. C. and Ashraf, I. A Systematic Review of Disaster Management Systems: Approaches, Challenges, and Future Directions. Land. (2023). 12(8).1514.

Kieu, Q. L. Nguyen, T. T. and Hoang, A. H. (2021). GIS And Remote Sensing: A Review of Applications to The Study of The Covid-19 Pandemic. Geography, Environment, Sustainability. (2021). 14(4).117-124.

Kingsley, J. Lawani, S. O. Esther, A. O. Ndiye, K. M. Sunday, O. J. and Penížek, V. Predictive mapping of soil properties for precision agriculture using geographic information system (GIS) based geostatistics models. Modern Applied Science. (2019). 13(10).60-77.

Lu, L. Li, S. Wu, R. and Shen, D. Study on the Scale Effect of Spatial Variation in Soil Salinity Based on Geostatistics: A Case Study of Yingdaya River Irrigation Area. Land. (2022). 11(10).1697.

Mathenge, M. Sonneveld, B. G. and Broerse, J. E. Application of GIS in Agriculture in Promoting Evidence-Informed Decision Making for Improving Agriculture Sustainability: A Systematic Review. Sustainability. (2022). 14(16).9974.

Molla, A. Zhang, W. Zuo, S. Ren, Y. and Han, J. A machine learning and geostatistical hybrid method to improve spatial prediction accuracy of soil potentially toxic elements. Stochastic Environmental Research and Risk Assessment. (2023). 37(2).681-696.

Mondal, B. P. Sekhon, B. S. Setia, R. K. and Sadhukhan, R. Geostatistical Assessment of Spatial Variability of Soil Organic Carbon Under Different Land Uses of Northwestern India. Agricultural Research. (2021). 10.407-416.

Mousavi, A. Karimi, A. Maleki, S. Safari, T. and Taghizadeh-Mehrjardi, R. Digital mapping of selected soil properties using machine learning and geostatistical techniques in Mashhad plain, Northeastern Iran. Environmental Earth Sciences. (2023). 82(9).234.

Nagaraj, S. Pandian, P. S. Mary, P. C. N. Geetha, R. and Gurusamy, A. Assessing spatial variability of soil and drawing location-specific management zones for coastal saline soils in Ramanathapuram District, Tamil Nadu. Journal of Applied and

Natural Science. (2023). 15(1).242-251.

Openshaw, S. and Clarke, G. (2019): Developing spatial analysis functions relevant to GIS environments. In Spatial analytical perspectives on GIS, Routledge, London.

Othmani, O. Khanchoul, K. Boubehziz, S. Bouguerra, H. Benslama, A. and Navarro-Pedreño, J. Spatial Variability of Soil Erodibility at the Rhirane Catchment Using Geostatistical Analysis. Soil Systems. (2023). 7(2).32.

Parker, D. (2023): Innovations in GIS. CRC Press. London.

Rajalakshimi, P. Mahendran, P.P. Mary, P.C.N. Ramachandran, J. Kannan, P. ChelviRamessh, and Selvam, S. Spatial Analysis of Soil Texture using GIS-based Geostatistics Models and Influence of Soil Texture on Soil Hydraulic Conductivity in Melur Block of Madurai District, Tamil Nadu. Agricultural Science Digest. (2023). 1-6.

Reza, S. K. Baruah, U. Sarkar, D. and Singh, S. K. Spatial variability of soil properties using geostatistical method: a case study of lower Brahmaputra plains, India. Arabian Journal of Geosciences. (2016). 9.1-8.

Singh, N. K. and Nathawat, M. S. Quantifying the characteristics and types of urban growth of Varanasi city using multi-temporal remotely sensed data and geospatial techniques. Transactions. (2023). 45(1).27.

Singh, S. and Sarma, K. Mapping Surface Soil Characteristics of Barren Land by Using Geospatial Technology in NCT of Delhi. Environment & We an International Journal of Science & Technology. (2020). 15(1).15-27.

Singh, S. and Sarma, K. Spatial variability of soil parameters: A geostatistical approach. International Journal of Geography Geology and Environment. (2023). 5(2).11-16.

Wang, F. and Liu, L. (2023): Computational Methods and GIS Applications in Social Science. CRC Press. Boca Raton.

Wang, X. Remote sensing applications to climate change. Remote Sensing. (2023). 15(3).747.

Xu, H. and Zhang, C. Development and applications of GIS-based spatial analysis in environmental geochemistry in the big data era. Environmental Geochemistry and Health.(2023). 45(4).1079-1090.

Zakeri, F. and Mariethoz, G. A review of geostatistical simulation models applied to satellite remote sensing: Methods and applications. Remote Sensing of Environment.(2021). 259. 112381.

Zhao, Z. Liu, G. Liu, Q. Huang, C. Li, H. and Wu, C. (2018). Distribution characteristics and seasonal variation of soil nutrients in the Mun River Basin, Thailand. International journal of environmental research and public health. (2018). 15(9).1818.

How to cite this article:

Singh, S., Sarma, K. Exploring Soil Spatial Variability with GIS, Remote Sensing, and Geostatistical Approach. Journal of Soil, Plant and Environment (2023); 2(1)-pp; 79-99.